Online News Popularity Prediction

Authors

Keerthi Rallapalli

KRallapalli@clarku.edu

Maitri Kotak

mkotak@clarku.edu

Shashank Gupta

shgupta@clarku.edu

Shivika Gupta

shigupta@clarku.edu

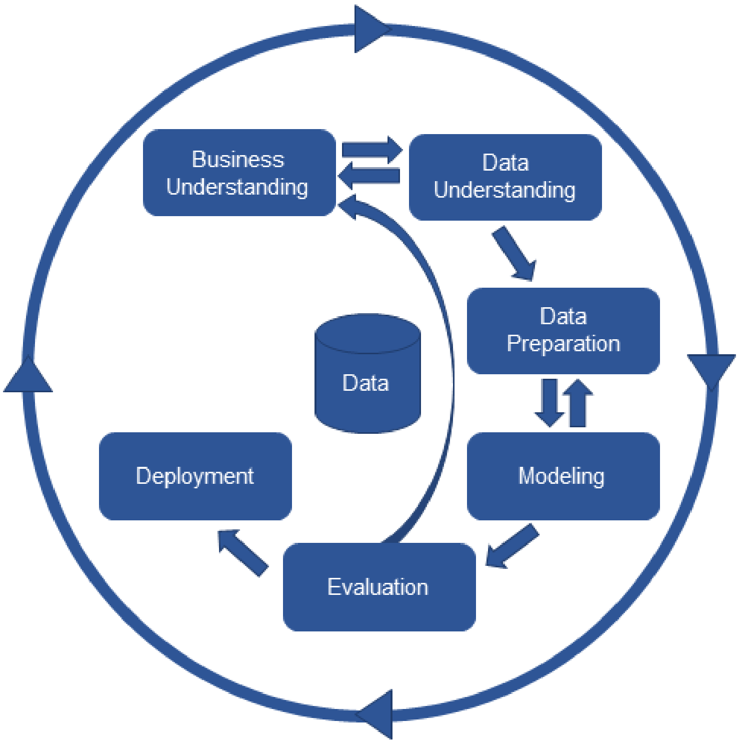
Abstract:

Online news popularity prediction aims to predict the future popularity of news article prior. estimating the number of shares. Popularity prediction is a challenging task due to various issues including difficulty to measure the quality of content and relevance of content to users; prediction difficulty of complex online interactions and information cascades, inaccessibility of context outside the web, local and geographic conditions, and social network properties. This paper focuses on popularity prediction of online news by predicting whether users an article is popular or not based on the number of shares. This paper proposes the popularity prediction using features that are known before publication of articles. The proposed model shows around 3.83% improvement between two different algorithms. This model also indicates that features extracted from articles keywords publication day, and the data channel are highly influential for popularity prediction.

1. Introduction

Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning, that analyse current and historical facts to make predictions about future or otherwise unknown events. In this project, Mashable Inc is the digital media website for which the predictive algorithms are being used. The data set summarizes a heterogeneous set of features about numerous articles published by Mashable in a period of two years. There are 61 attributes out of which 58 are predictive attributes, 2 non-predictive and 1 goal field. Our main aim is to conduct research on whether the given news article item is popular or not. The given dataset contains number of shares as the predictor variable.

CRISP-DM, which stands for Cross-Industry Standard Process for Data Mining, is an industry-proven way to guide the data mining efforts. By following the methodology of the CRISP-DM process, every step of the project is defined. The following diagram explains it more:



To explain the process further, Business Understanding involves the study of the business domain. In this case, we study the digital media industry where Mashable Inc is one of the key players and the understand the problem at hand. Data Understanding deals with the study of the various independent variables in the dataset and also the dependent variable. Data cleaning is followed by data understanding. Here, we deal with the missing values, outliers and NA values. Because we want to understand each of the variables, the correlations among them and also their relationship with the predictor variable, we proceed with Exploratory Data Analysis. As part of this, we do univariate, bivariate and multivariate analyses. After the in-depth analysis of each of the variables, we made use of IVA or Information Value Analysis. This was used to determine which out of all the 59 variables produce more variation in the predictor variable – number of shares. This in turn, helped in narrowing the number of independent variables on the number of shares. This also marks the final step in which we prepare the data for the modelling step.

Modelling is the step wherein the technique used for modelling is identified and stated along with any assumptions if required. In this case, Random Forest and Logistic Regression algorithms are identified. This is also the stage where the dataset is divided into testing and training datasets. The model is built and assessed on the training dataset first. The parameters are well defined and revised according to the need after the assessment procedure.

Evaluation is the stage where data mining results are assessed, and approved models are finalized for implementation further. This is the stage where the entire process is summarized and those activities which were missed are highlighted for use in the future. All those activities that are approved to be used further are also stated to finish finalizing the exact modelling procedure. The decision on how to proceed is stated along with the rationale.

Deployment is the final step in the process wherein we summarize the strategy to deploy the project and how to perform the steps in order to implement in real-world.

2. Methods and Analysis

Data has been imported from the UCI Machine Learning Repository. The dataset summarized a heterogeneous set of featured articles published by Mashable in a period of two years.

There are 61 variables considered for the purpose of study, out of which 58 are predictor variables and 2 are non-predictive variables. The variable ‘shares’ is the target variable, which has been converted to a categorical variable ‘popularity’, based on the number of shares.

Variable Description:

0. url: URL of the article (non-predictive)

1. timedelta: Days between the article publication and the dataset acquisition (non-predictive)

2. n_tokens_title: Number of words in the title

3. n_tokens_content: Number of words in the content

4. n_unique_tokens: Rate of unique words in the content

5. n_non_stop_words: Rate of non-stop words in the content

6. n_non_stop_unique_tokens: Rate of unique non-stop words in the content

7. num_hrefs: Number of links

8. num_self_hrefs: Number of links to other articles published by Mashable

9. num_imgs: Number of images

10. num_videos: Number of videos

11. average_token_length: Average length of the words in the content

12. num_keywords: Number of keywords in the metadata

13. data_channel_is_lifestyle: Is data channel 'Lifestyle'?

14. data_channel_is_entertainment: Is data channel 'Entertainment'?

15. data_channel_is_bus: Is data channel 'Business'?

16. data_channel_is_socmed: Is data channel 'Social Media'?

17. data_channel_is_tech: Is data channel 'Tech'?

18. data_channel_is_world: Is data channel 'World'?

19. kw_min_min: Worst keyword (min. shares)

20. kw_max_min: Worst keyword (max. shares)

21. kw_avg_min: Worst keyword (avg. shares)

22. kw_min_max: Best keyword (min. shares)

23. kw_max_max: Best keyword (max. shares)

24. kw_avg_max: Best keyword (avg. shares)

25. kw_min_avg: Avg. keyword (min. shares)

26. kw_max_avg: Avg. keyword (max. shares)

27. kw_avg_avg: Avg. keyword (avg. shares)

28. self_reference_min_shares: Min. shares of referenced articles in Mashable

29. self_reference_max_shares: Max. shares of referenced articles in Mashable

30. self_reference_avg_sharess: Avg. shares of referenced articles in Mashable

31. weekday_is_monday: Was the article published on a Monday?

32. weekday_is_tuesday: Was the article published on a Tuesday?

33. weekday_is_wednesday: Was the article published on a Wednesday?

34. weekday_is_thursday: Was the article published on a Thursday?

35. weekday_is_friday: Was the article published on a Friday?

36. weekday_is_saturday: Was the article published on a Saturday?

37. weekday_is_sunday: Was the article published on a Sunday?

38. is_weekend: Was the article published on the weekend?

39. LDA_00: Closeness to LDA topic 0

40. LDA_01: Closeness to LDA topic 1

41. LDA_02: Closeness to LDA topic 2

42. LDA_03: Closeness to LDA topic 3

43. LDA_04: Closeness to LDA topic 4

44. global_subjectivity: Text subjectivity

45. global_sentiment_polarity: Text sentiment polarity

46. global_rate_positive_words: Rate of positive words in the content

47. global_rate_negative_words: Rate of negative words in the content

48. rate_positive_words: Rate of positive words among non-neutral tokens

49. rate_negative_words: Rate of negative words among non-neutral tokens

50. avg_positive_polarity: Avg. polarity of positive words

51. min_positive_polarity: Min. polarity of positive words

52. max_positive_polarity: Max. polarity of positive words

53. avg_negative_polarity: Avg. polarity of negative words

54. min_negative_polarity: Min. polarity of negative words

55. max_negative_polarity: Max. polarity of negative words

56. title_subjectivity: Title subjectivity

57. title_sentiment_polarity: Title polarity

58. abs_title_subjectivity: Absolute subjectivity level

59. abs_title_sentiment_polarity: Absolute polarity level

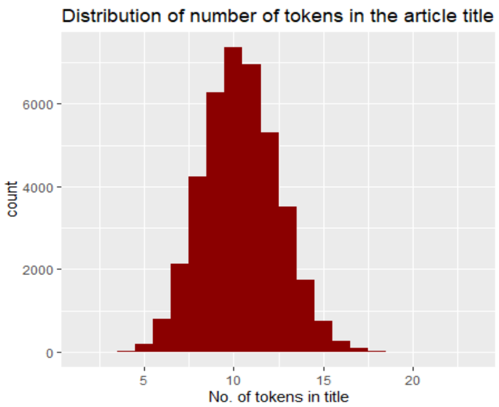
60. shares: Number of shares (target)

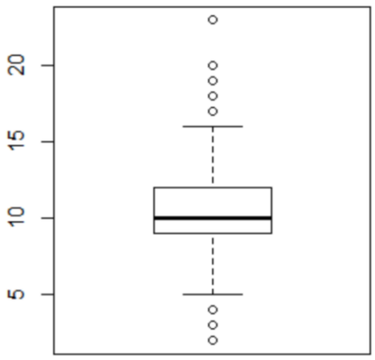
The data has been prepared using univariate analysis, bivariate analysis and information value analysis.

1) **Univariate Analysis -** Univariate analysis has been done using graphical analysis and boxplot. The analysis has been performed in order to examine the distribution and outliers of all the numerical and categorical variables. Since it is not practically possible to describe the nature of each of the 61 variables, we considered explaining some of the variables that turned out to be significantly affecting the output through Exploratory Data Analysis.

Numerical variables:

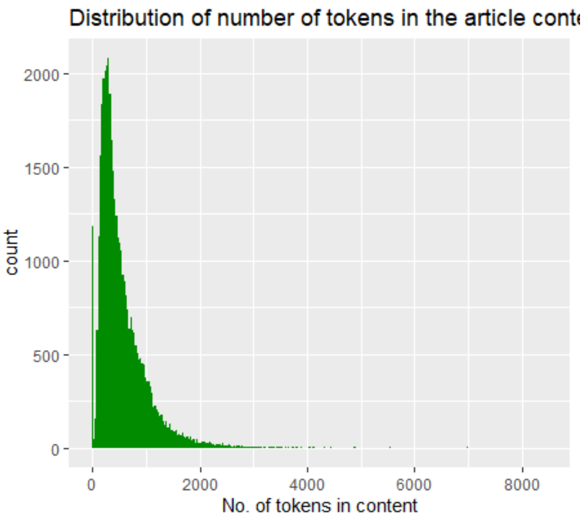
- Examining the number of tokens in the articles:

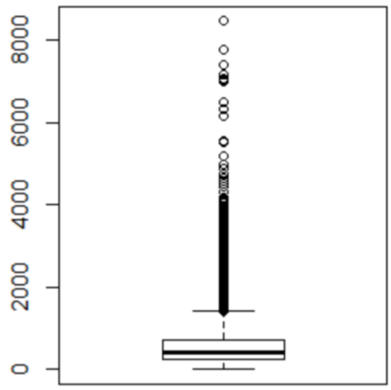




After careful observation of the distribution we found that the distribution is normal. There a total of 156 outliers detected which comprised of approximately 0.4% of the whole dataset which is insignificant.

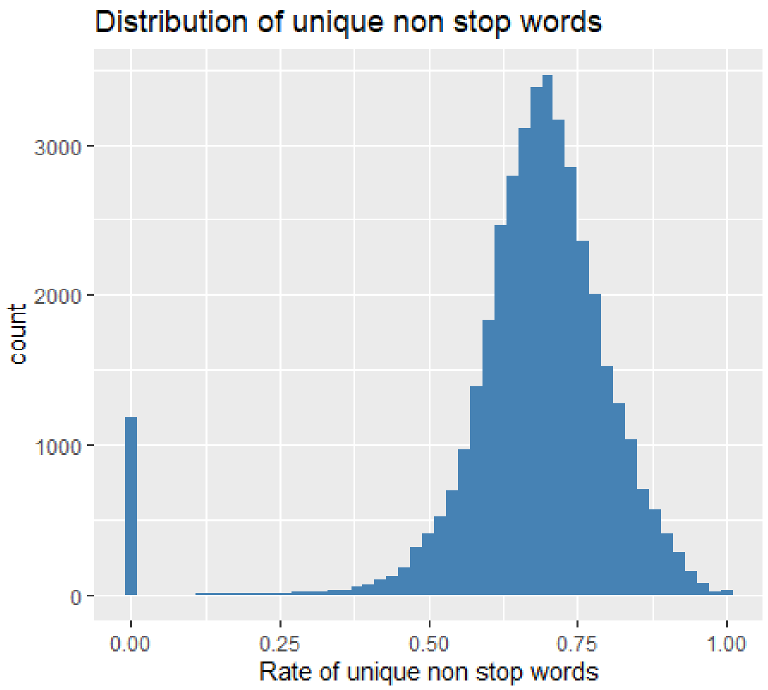
- Examining the number of tokens in the content of the article:





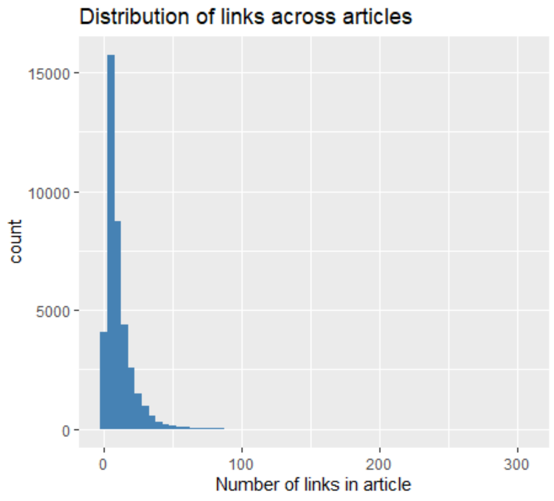
After careful observation of the distribution we found that the distribution is skewed. There were a total of 1933 outliers detected which comprised of approximately 5% of the whole dataset.

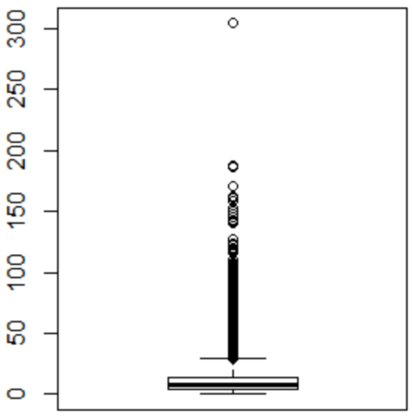
- Examining the number of non-stop unique tokens distribution in the articles



After careful observation of the distribution we found that the distribution is nearly normal with a few data points lying on the left of the normal curve which is slightly skewed. There were no outliers detected.

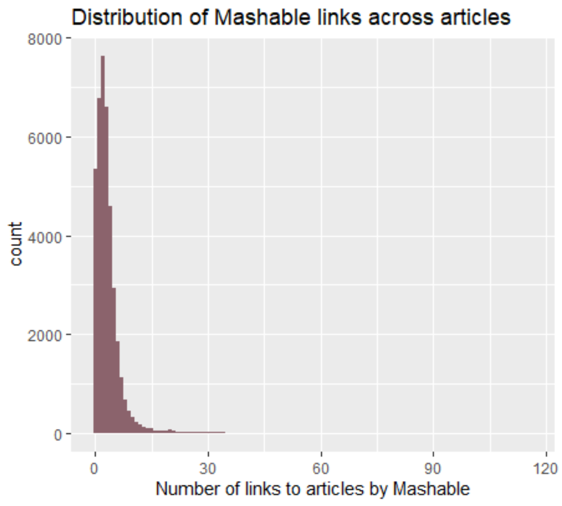
- Examining the distribution of number of links in articles

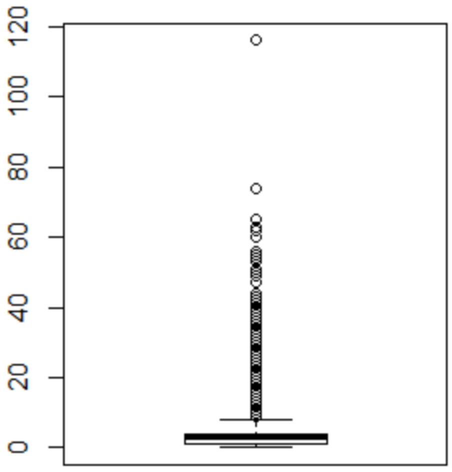




The distribution of links across articles is right skewed, with nearly 5.5% of outliers.

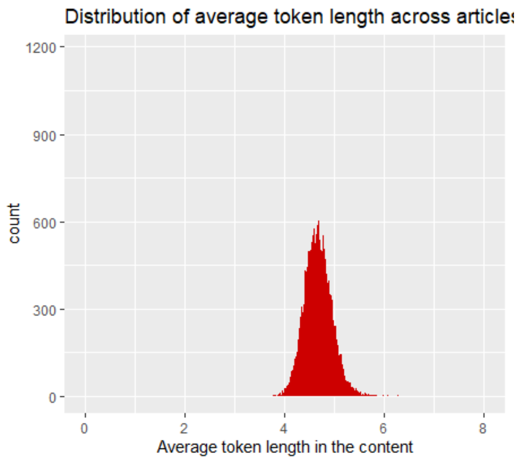
- Examining the distribution of links in articles to the articles published by Mashable

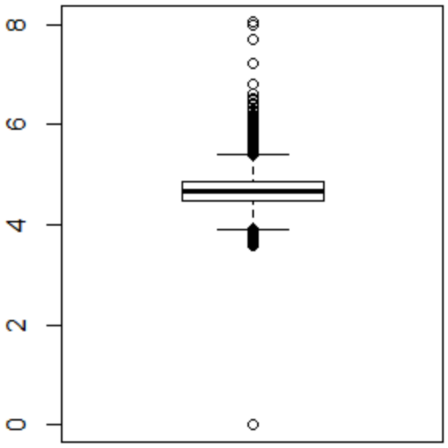




After careful observation of the distribution we found that the distribution is highly skewed. There were a total of 2090 outliers detected which comprised of approximately 5.27% of the whole dataset.

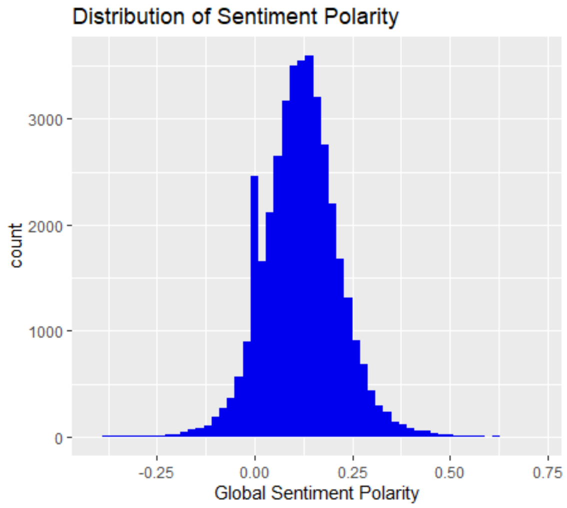
- Examining average length of words in the content

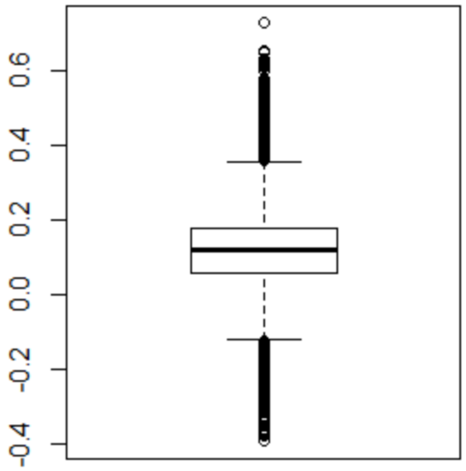




After careful observation of the distribution we found that the distribution is nearly normally distributed. Most of the articles have an average length between 4 and 6. There were a total of 1681 outliers detected which comprised of approximately 4.24% of the whole dataset.

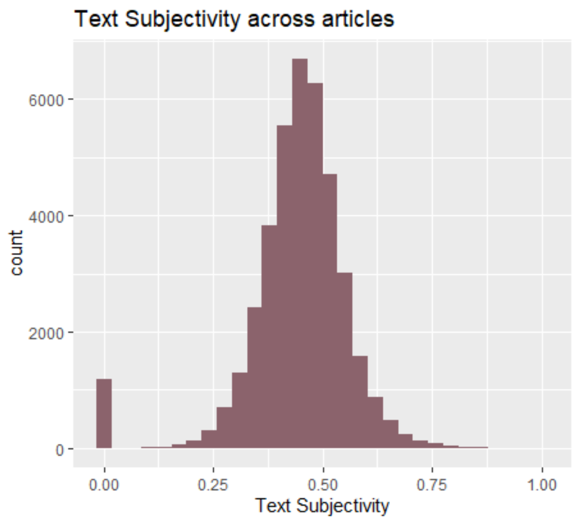
- Examining the text sentiment polarity

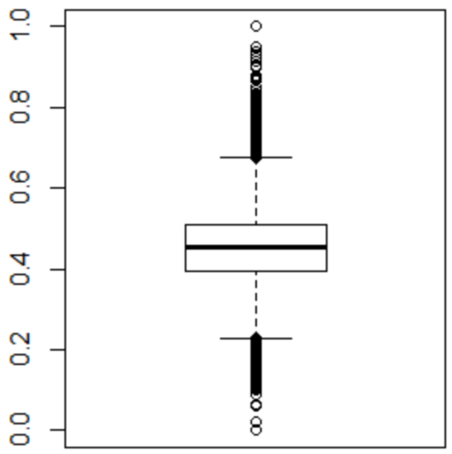




After careful observation of the distribution we found that the distribution is normal. There were a total of 825 outliers detected which comprised of approximately 2.08% of the whole dataset.

- Examining the text subjectivity



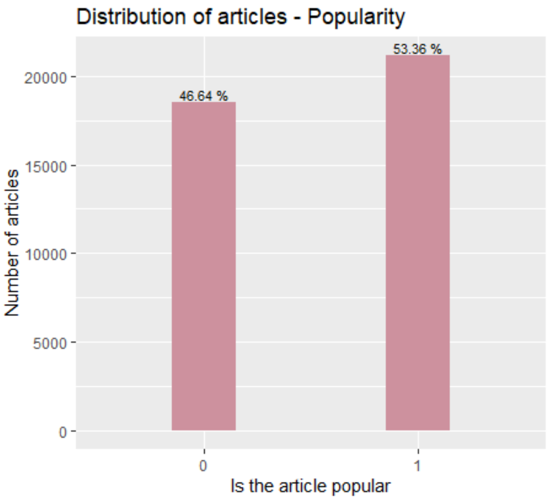


After careful observation of the distribution we found that the distribution is normal. There were a total of 1912 outliers detected which comprised of approximately 4.82% of the whole dataset.

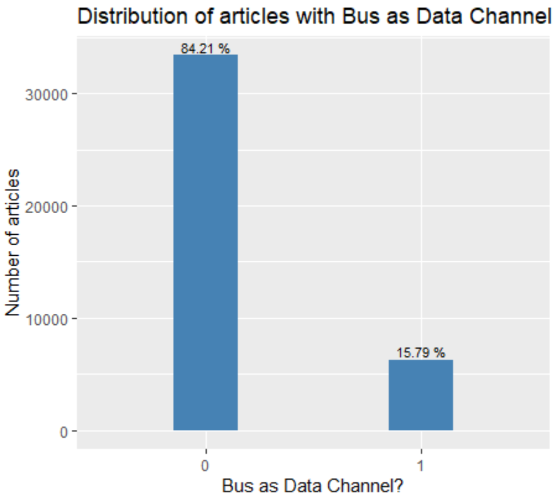
Categorical Variables

To define the distribution of the categorical variable shares (target variable) we took the median of this variable and defined it as the threshold between popular or not popular as the proportion of the articles were 53.36% & 46.64% respectively. When we tried defining threshold using other statistical metrics like mean and quartiles, we didn’t find the balanced proportion of the number of articles between popular and not popular.

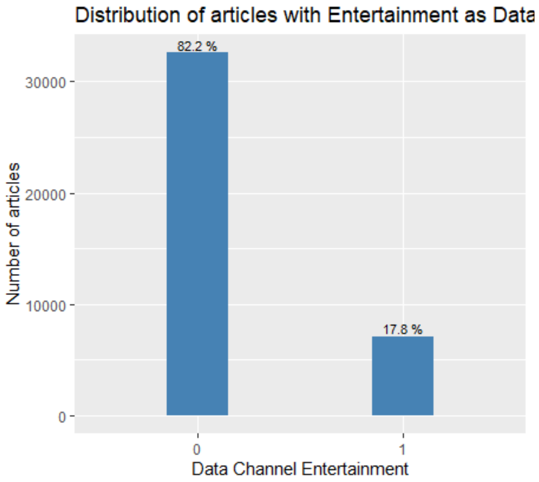
Examining the popularity variable



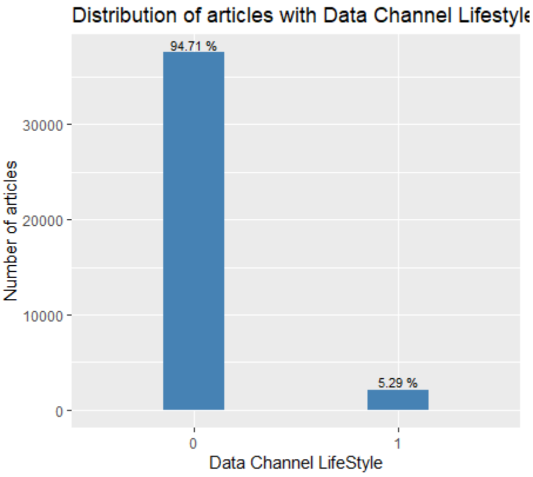
Examining the category Business



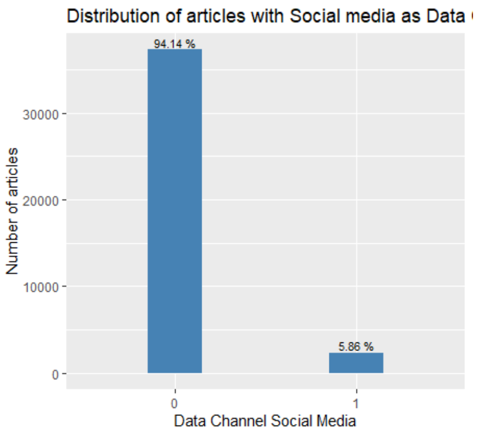
Examining the category Entertainment



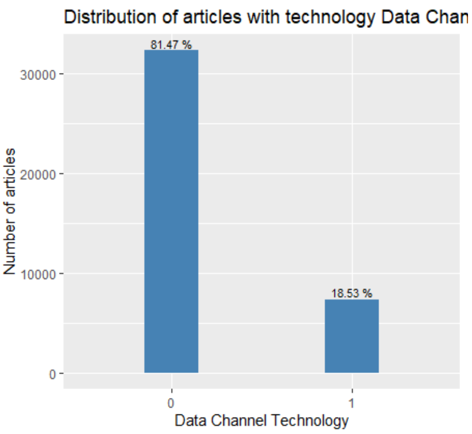
Examining the category Lifestyle



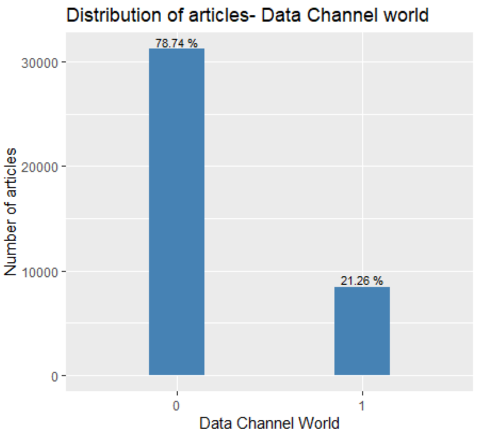
Examining the category Social Media



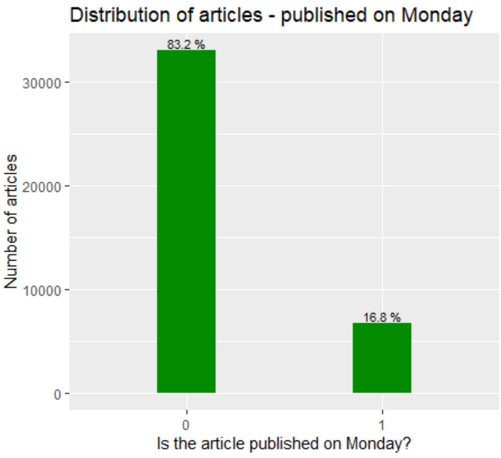
Examining the category Technology



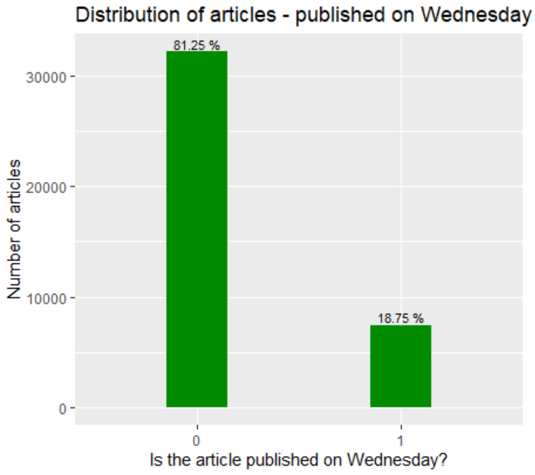
Examining the category World



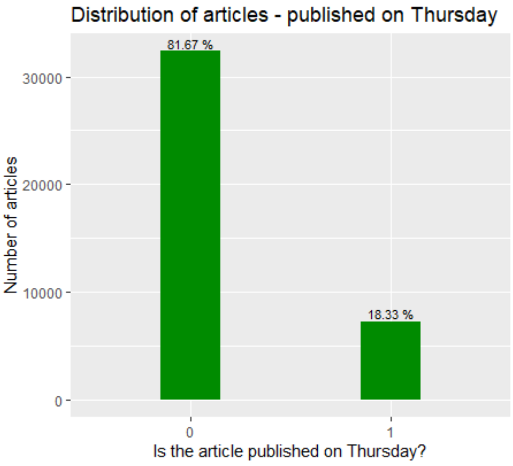
Examining the weekday Monday



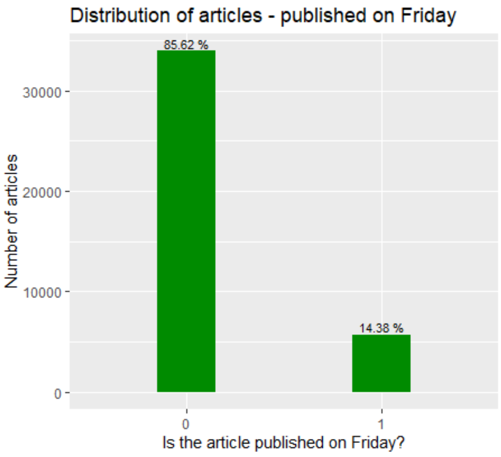
Examining the weekday Tuesday



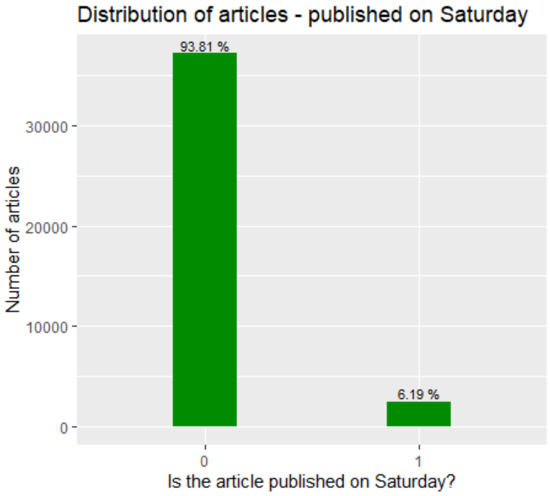
Examining the weekday Thursday



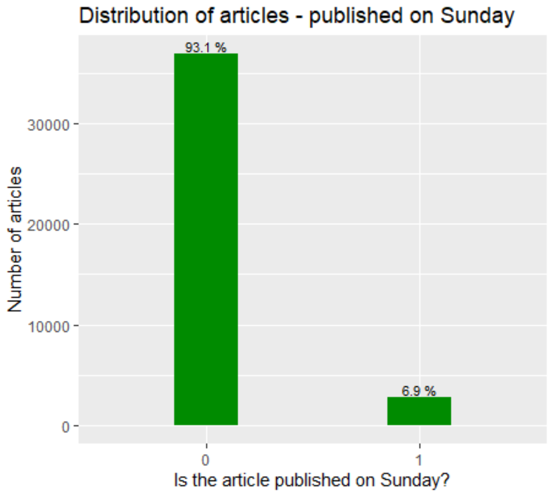
Examining the weekday Friday



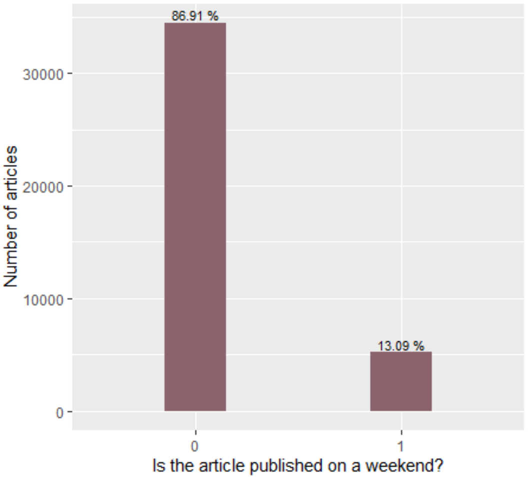
Examining the weekday Saturday



Examining the weekday Sunday

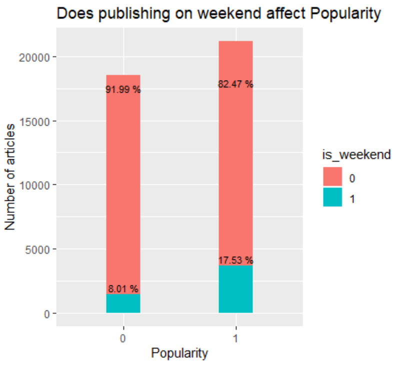


Examining the variable Weekend

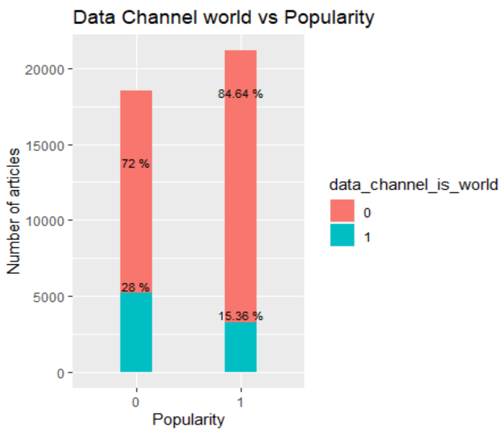


2) Bivariate Analysis -

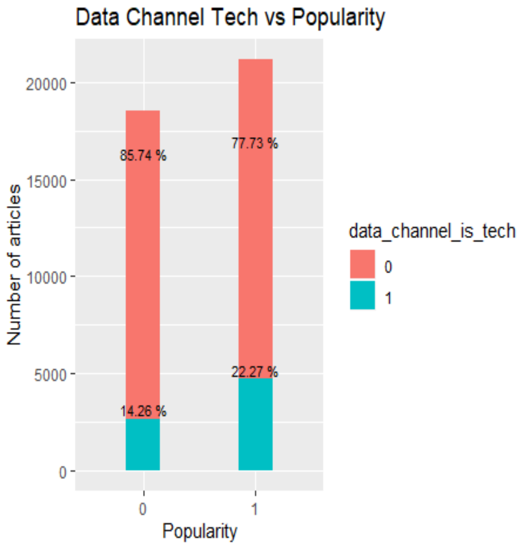
In Bivariate analysis, the variables are analysed and compared with the target variable popularity.



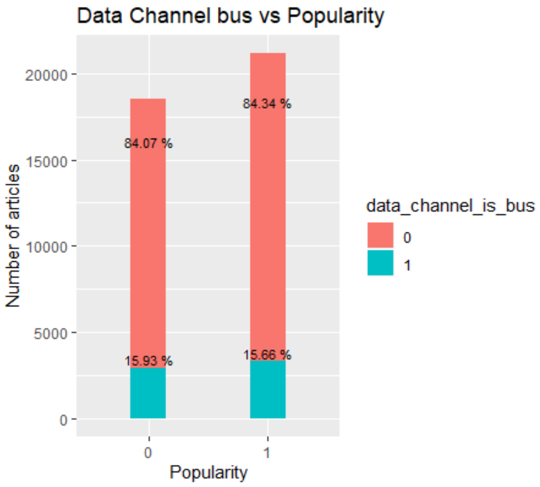
In this graph we can see that if an article is published on weekend then popularity is 17.53% as compared to the other days, however, in the case of non-popular articles, only 8.01% are published on weekends as compared to other days.



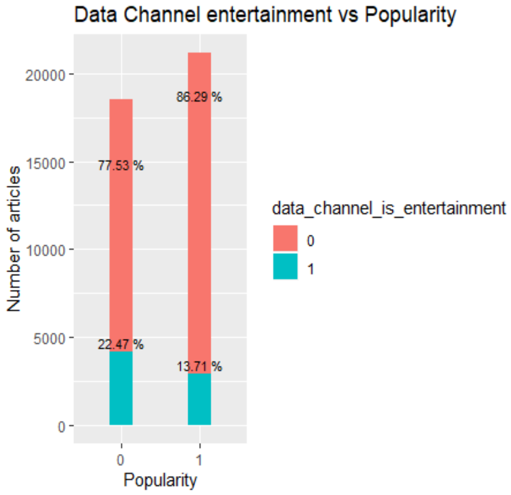
This graph shows popularity based on channel world, i.e. if the channel chosen is world then popular articles are 15.36% as compared to other channels. Also, if the Channel is world then the articles not popular are 28% so a greater proportion of articles are unpopular in this channel.



This graph shows popularity based on channel technology, i.e. if the channel chosen is technology, then popular articles are 22.27% as compared to other channels. Also, if the Channel is technology then the articles that are not popular are 14.26% so a greater proportion of articles are popular in this channel.

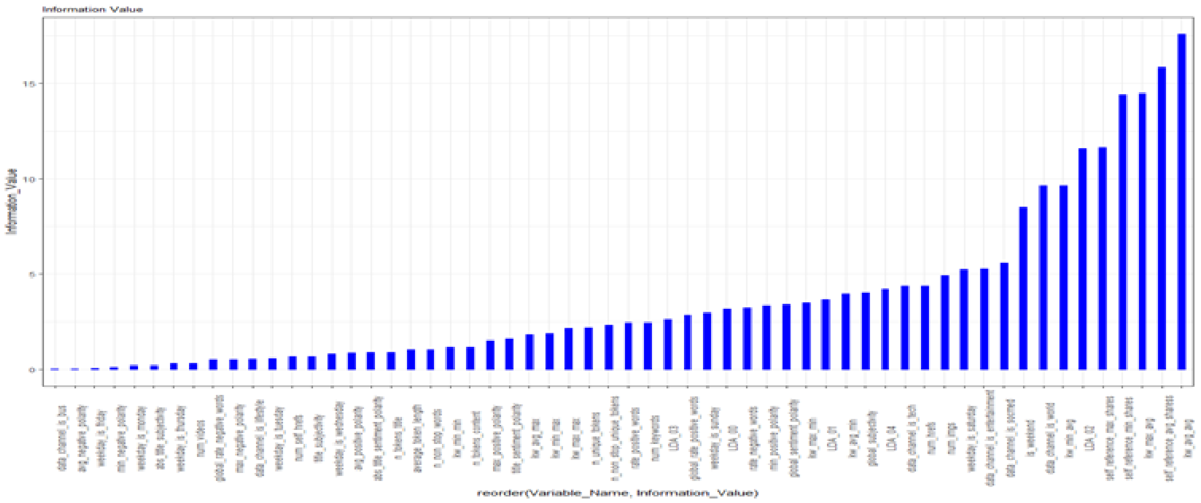


This graph shows popularity based on channel business i.e. if the channel chosen is business then popular articles are 15.66% as compared to other channels. Also, if the Channel is business then the articles that are not popular are 15.93% so the proportion of articles that are popular and unpopular in this channel are almost equal.



This graph shows popularity based on channel entertainment, i.e. if the channel chosen is entertainment then popular articles are 13.71% as compared to other channels. Also, if the Channel is entertainment then the articles not popular are 22.47% so a greater proportion of articles are unpopular in this channel.

3) Information Value Analysis -



We are performing Information Value Analysis (I.V.) since our dataset has too many variables that can complicate the model and lead to overfitting. Using Information Value Analysis, we end up selecting 25 Variables that have a significant effect over the target variable since the variation of these variables is high. For this, we set a threshold of I.V. >3 because variables with smaller I.V. i.e. less than 3 can be ignored due to their insignificant effect over the target variable.

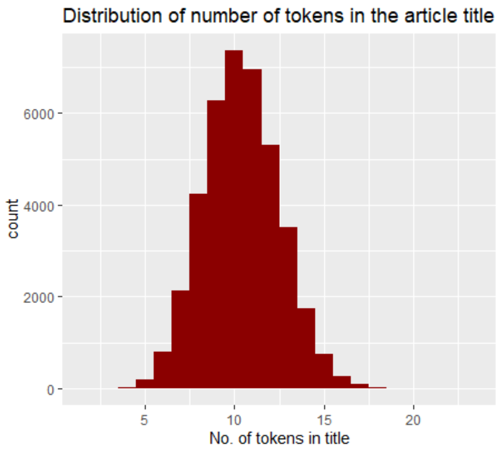
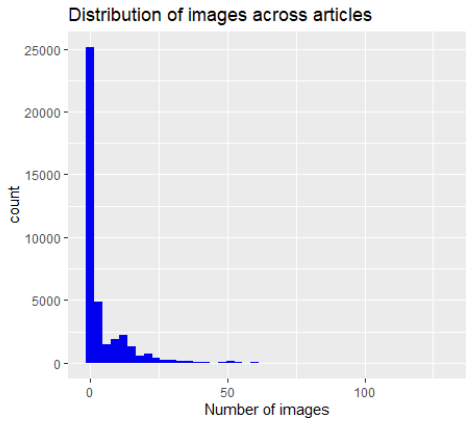
Data cleaning was not a tedious task as we got a cleaned data to analyse. We checked the null or NA values using the standard data cleaning method but didn't find any NA or empty value. There was a column ' that had to have to values between 0 to 1 and had 7 as a value

In order to check if there were any outliers in the variables, box plots have been graphed and we did not remove any variable because the outliers constituted a large proportion of the population and could significantly affect the results. There were no missing values that were needed to be replaced hence we didn't impute any missing values.

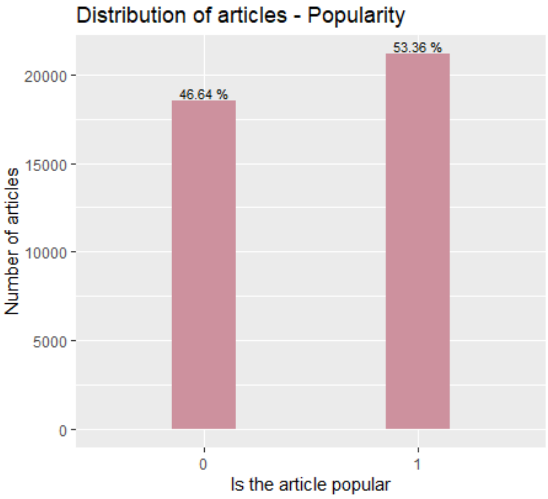
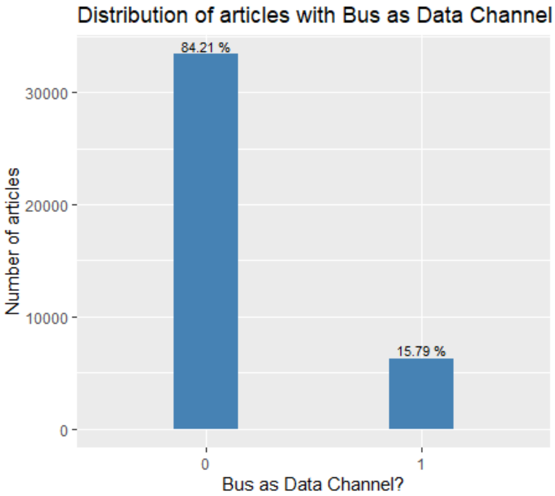
3. Results

After completing the analysis, we summarized our findings after which we would apply algorithms like Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM).

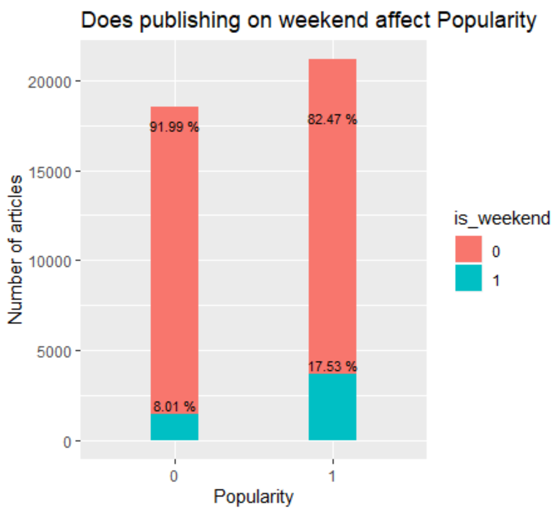
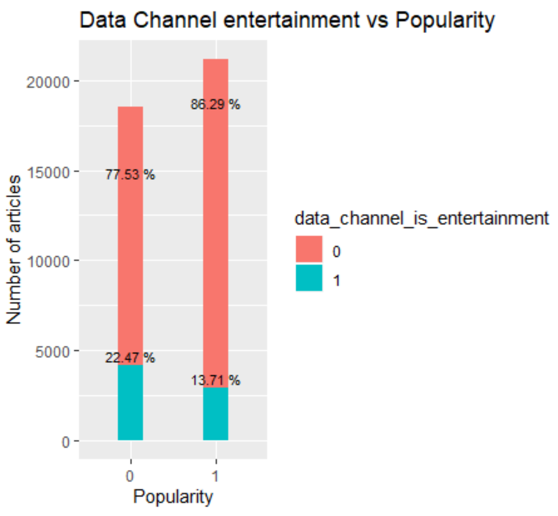
As we performed univariate, bivariate and information value analysis, we found very interesting metrics that should be considered for data pre-processing.
Univariate analysis - In this analysis, we can easily see the distribution type of the numerical variables where the variables performed from normal distribution to being rightly skewed. The numerical variables distribution of number of tokens in the title, unique non-stop words, average word length, text subjectivity are almost normally distributed while the variables like number of tokens in the content, the number of videos and images, the number of links in an article have skewed distribution.

	
Example of numerical univariate normal distribution	**Example of numerical univariate skewed distribution**

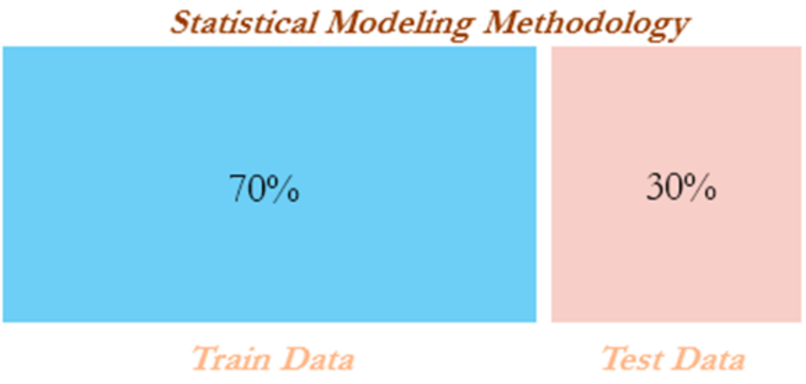
The categorical variables in the univariate analysis shows the Yes/No about an article. The categorical variables included majorly the data channels (category in which an article published) and the weekday when the article was published and the popularity of an article (based on the threshold).

	
Example of categorical popularity univariate analysis	**Example of categorical univariate analysis**

Bivariate analysis - In the bivariate analysis, we compared the categorical variables with the popularity variable to see the relation between them and how exactly are they affecting the popularity of an article. We compared the variables like the weekday on which an article is published, category in which an article is published with the popularity of an article.

	
Example of comparison of categorical variable with popularity	**Example of comparison of categorical variable with popularity**

We divided the training and test dataset in 7:3 ratio to find the accuracy of the model.



We applied three algorithms - Logistic Regression, Random Forest Classifier and Support Vector Machine – that helped in building models with the train dataset. We were unable to implement SVM due to computational power constrains (limited RAM).

1. Logistic Regression:

   It is used since the dependent variable is categorical and using a standard linear model would be inappropriate since the dependent variable can only be 0 and 1 i.e. non-popular and popular respectively. In this model instead of predicting the value of the dependent variable, we predict the probability of the dependent variable equal to 1. The model has been executed with nine iterations in order to remove any variables that showed signs of multicollinearity or variables that do not significantly impact the target variable. To remove these, we used a threshold of 7 for Variation Inflation Factor and p-value of 0.05.

2. Random Forest Classifier: Random forest classifier is an ensemble of trees that can improve the prediction accuracy of a model, through numerous iterations. The principle of cross validation is in-built in the model, since there are multiple values of hyperparameters that are tuned during the process. It makes use of numerous decision tree classifiers on various sub-samples of the dataset and averages all of the best results.

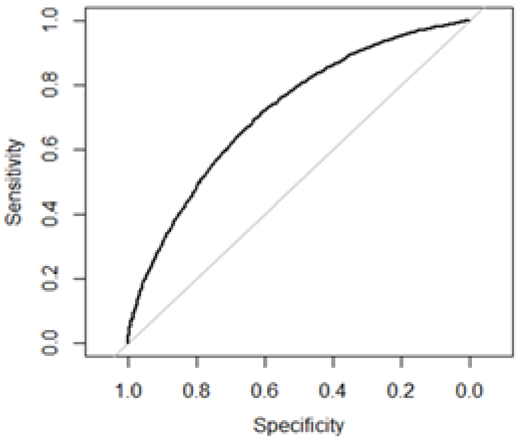
Summary of Accuracy of the models:

Algorithm	**Accuracy**
Logistic Regression	63.41%
Random Forest Classifier	71.93%

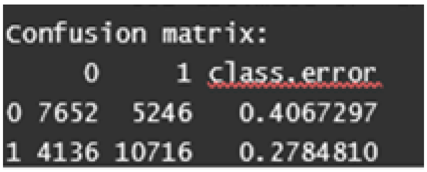
4. Discussion and Conclusion

ROC shows how better the model distinguishes between different classes i.e. popular and non-popular in this case. The receiver operating curve is a plot of the true positive rate against the false-positive rate. .Prediction result or any point in the confusion matrix represents one space in ROC.

1. Area under ROC curve for random forest classifier is 71.93%

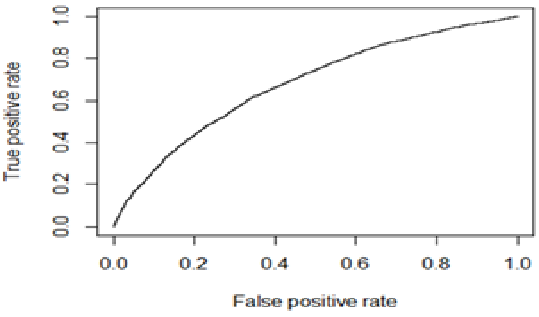


Confusion matrix contains information about actual data and the predicted data using the classification system. This is the confusion matrix for random forest model which shows that the false popular articles predicted are 5246 and false non-popular articles predicted by this model are 4136 whereas true popular and non-popular articles predicted by this model are 10716 and 7652 respectively.



```
Confusion matrix:
     0     1 class.error
0 7652  5246   0.4067297
1 4136 10716   0.2784810
```

2. Area Under the ROC curve for Logistic Regression is 68.10%



Sensitivity : 0.7221

Specificity : 0.5317

Accuracy : 0.6326

After careful consideration we found the following variables that affect the popularity of a news article:

1. An article published during the weekend has a better chance of receiving more shares and becoming popular.
2. If the article is published under the category entertainment, social media, and technology then it has the better probability of getting more shares and becoming popular.
3. Articles with a higher number of images shows a tendency to become more popular and result in more shares.
4. Articles that has more than the average number of keywords to become popular and result in more shares.
5. Global text subjectivity should be always taken into consideration while writing an article.
6. Negative keywords should be less as it might result in low popularity of an article.

Limitations:

As we come to the end of our paper, we would like to highlight a few of the limitations of this paper. The dataset that we chose has too many variables due to which we faced computational power issue while implementing support vector machine algorithm. Further, due to multiple variables we faced overfitting problem due to which the accuracy of logistic regression algorithm was quite low even after solving the multicollinearity issue. The specificity obtained for the Logistic Regression is low (53%). There is scope for further improvement of the categorization of false positives.

Business Recommendations:

Based on our research using the data analytics we are recommending few suggestions for online news business:

1. Articles published on weekend have a higher likelihood of becoming popular.
2. Articles published under the category of technology or entertainment or social media are more likely to receive higher number of shares.
3. More the number of links and images, more is the likelihood that the article becomes popular.